DIALPFN: ASSUMPTION-CONDITIONED IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In time-to-event studies, some individuals leave follow-up before the event occurs, leaving its timing unknown. Predictions then depend on assumptions about how event times relate to when follow-up ends, a relationship that the observed records alone generally cannot determine. Analysts therefore need to understand how their predictions change under alternative assumptions, not just obtain one answer under a fixed assumption. We introduce DialPFN, a prior-fitted network that accepts an observed dataset, an individual's features, and a stated dependence assumption, and returns an event-time distribution under that assumption. Synthetic studies provide complete target times for supervision, while the network receives only the corresponding partially observed table and a stated dependence assumption. At inference, changing this input produces a sensitivity analysis for each query without updating model weights. A frozen-backbone quantile head and reference-sample bagging produce individual predictive distributions. For the reference predictor, supplying the generating assumption reduces strong-dependence survival-curve error by 48.3 to 54.7% across five reserved simulation banks. Ordinary prediction is preserved across 17 real datasets, and a separate Melanoma evaluation supports stable individual sensitivity curves. All six training runs pass seven grouped evaluation checks. DialPFN makes individual predictions conditional on explicit assumptions, allowing analysts to examine how their conclusions change across plausible alternatives.

1 INTRODUCTION

When follow-up ends early, the reason it ends can matter for predicting what happens next. This uncertainty has consequences in both medical and economic settings.

Consider a cancer study in which patients stop attending before progression is recorded. Among patients with similar recorded characteristics, earlier exit might accompany earlier disease progression, or the two times might be unrelated (Gupta et al., 2025). These assumptions lead to different interpretations of the missing outcomes and can change a new patient's predicted probability of remaining progression-free at five years. The analyst needs to know how much that probability depends on the assumed relationship.

An employment survey presents a similarly consequential question (Van Den Berg & Lindeboom, 1998). Suppose job seekers leave follow-up while still unemployed, and those who leave sooner tend to find work sooner afterward. Their later job dates remain unrecorded. Assuming unrelated survey exit can then distort the predicted probability that a job seeker remains unemployed after six months. For income-support and employment-service planning, it matters whether a conclusion holds across plausible explanations of those missing records.

Related concerns arise when preventive maintenance removes equipment before failure (Bedford, 2004), when radio-collared wildlife disappears and its unknown fate may conceal poaching (Liberg et al., 2011), and when participants in homelessness studies are lost to follow-up before reaching stable-housing milestones (Braciszewski et al., 2016). Assumptions about these unobserved outcomes can affect maintenance, conservation, and housing-support decisions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The observed records alone generally cannot determine the required relationship. Different joint distributions of event and stopping times can produce the same distribution of observed times and labels (Tsiatis, 1975). Collecting more records of the same kind can estimate that observed distribution more precisely while leaving this ambiguity unresolved. Additional model restrictions or external information can narrow the possibilities. Sensitivity analysis addresses this uncertainty by holding the observed evidence fixed, varying the stated assumption, and examining how predictions change. Existing statistical methods already support this analysis, including copula models for dependent censoring (Yeh et al., 2023). Meanwhile, survival prior-fitted networks (PFNs) learn reusable prediction procedures with conditionally independent censoring in their training setup (Qi et al., 2026; Böhm et al., 2026). Together, these lines of work motivate a prediction procedure that combines the explicit assumptions of statistical sensitivity analysis with the reusable inference of PFNs.

We introduce DialPFN, a prior-fitted network that makes unresolved dependence assumptions explicit inputs to individual event-time prediction. It accepts an observed reference table, an individual's features, and a stated dependence assumption, and returns an event-time distribution under that assumption. Holding the table and individual fixed while varying the assumption produces an individual sensitivity curve, without fitting model weights to the new study or retraining for each assumption. When the records distinguish several processes that stop target observation, each modeled process receives a coordinate of the same assumption input. Study exit and competing events therefore use one shared dependence-conditioning mechanism.

Simulation supplies the hidden-target supervision needed to learn this procedure. Following the PFN approach (Müller et al., 2021), we train across simulated studies with varied outcomes and dependence settings. The simulator retains complete event times, while the network receives incomplete reference records and query features. Complete query times serve as training answers. At inference, the learned weights remain fixed, and predictions are conditional on both the supplied assumption and the simulated training prior.

Our contributions are (1) a prediction task with explicit dependence inputs, (2) a simulation-trained implementation that returns individual distributions, and (3) an evaluation separating hidden-target accuracy, ordinary predictive fit, and sensitivity stability. For the reference predictor, supplying the generating assumption reduces strong-dependence curve error by 48.3 to 54.7% across five reserved simulation banks. Ordinary prediction is preserved across 17 real datasets relative to the earlier Gaussian-head predictor. All six training runs pass the complete set of seven checks. Together, these results evaluate whether the learned predictions use the assumption effectively while retaining useful prediction and uncertainty estimates.

2 Related work

Sensitivity analysis under dependent observation. Classical work establishes why dependence assumptions matter when only the first event or stopping time is observed: different joint-time distributions can yield the same observed records (Tsiatis, 1975). Copula methods make this dependence explicit and support survival prediction under alternative assumptions (Yeh et al., 2023). A copula specifies how event and stopping times are associated while separating that association from their individual distributions. This provides the statistical basis for our task: calculate predictions conditional on a stated relationship, then examine their sensitivity to that choice. The target also matters. In competing-event settings, net-survival interpretations depend on cause definitions (National Cancer Institute, n.d.); a latent event-time distribution and the probability of an observed event occurring first answer different questions. DialPFN adopts the latent-time prediction task and learns a procedure that can be reused across studies and dependence settings.

Learning reusable prediction from simulation. Simulation-based inference provides a way to learn such reusable procedures when a simulator supplies complete training answers. Papamakarios & Murray (2016) learn conditional density estimators from simulations for inference in generative models. PFNs apply a related learning principle to prediction: train across synthetic datasets so that a network approximates posterior prediction under a training prior (Müller et al., 2021). TabPFN demonstrates this approach for tabular classification and regression (Hollmann et al., 2025). PFNs establish the value of learning across tasks before observing the dataset of interest. For DialPFN, the additional requirement is to preserve an analyst's control over a specified part of the

data-generating model. The observed table supplies the evidence, while the dependence input selects the conditional prediction requested from the learned procedure.

Survival PFNs and dependence assumptions. Survival PFNs bring reusable prediction directly to incomplete event-time tables. SurvivalPFN learns in-context Bayesian survival prediction with conditionally independent censoring (Qi et al., 2026). SurvPFN uses non-informative censoring and a censored likelihood to train a distributional regression PFN (Böhm et al., 2026). Both provide a close foundation for our setting, but their training setups fix the relationship between target and censoring times through independence assumptions. DialPFN instead trains across dependence settings and receives the stated setting at inference. This extends the prediction task from answering under a fixed censoring assumption to producing a family of individual distributions as that assumption varies, with the observed table held fixed.

Sensitivity PFNs and the quantity being predicted. The closest connection between PFNs and sensitivity analysis is Javurek et al. (2026), who amortize causal sensitivity analysis. Their procedure estimates bounds on a causal effect across explanations allowed by a specified level of hidden confounding. DialPFN addresses a different question: under a specified dependence assumption and training prior, what event-time distribution should be predicted for an individual? For example, the output gives a probability of remaining progression-free at a chosen horizon under that assumption. The causal PFN also represents uncertainty about its bounds; our prediction intervals describe variation in an event time within a stated model. Bounds across compatible explanations and a distribution conditional on one assumption therefore have distinct interpretations. A sweep of DialPFN inputs traces conditional predictions; establishing causal bounds would require a separate identification argument.

Established components for distributional prediction. Quantile regression provides a standard way to learn multiple parts of an outcome distribution (Koenker & Bassett Jr, 1978), while bagging aggregates predictions across resampled datasets (Breiman, 1996) and has been used in survival learning (Hothorn et al., 2004). TabPFN Unleashed also uses reference resampling and per-query probability averaging (Liu & Ye, 2025). DialPFN builds on these components through a quantile head and reference-sample bagging, keeping each individual's predictions separate. They support the implementation of the central task: reusable individual event-time prediction conditional on an explicit dependence assumption.

3 Prediction under Explicit Dependence Assumptions

Observed records and hidden targets. Let T denote the target event time and C_j, $j = 1, \ldots, K$, the times of K labeled processes that can end its observation. The formulation allows multiple stopping processes; the implementation evaluated here uses $K = 2$. Supporting additional processes would require extending the model inputs and simulator and retraining the model, which is outside the scope of the present experiments. We include the target event in the same index set by defining $C_0 = T$; the stopping processes retain indices $1, \ldots, K$. A reference record contains features x, the first recorded time Y, and its label J:

$$Y = \min_{0 \le j \le K} C_j, \qquad J = \operatorname*{argmin}_{0 \le j \le K} C_j. \tag{1}$$

Here $J = 0$ identifies the target event and $J = j \ge 1$ identifies stopping process j. The reference table is $\mathcal{D} = \{(x_i, Y_i, J_i)\}_{i=1}^{n}$. When a stopping process occurs first, the record tells us that $T > Y$ but leaves T unknown. A new query supplies features x_*, with its outcome withheld.

The conditional prediction task. The analyst supplies an assumption vector $\boldsymbol{a} = (a_1, \ldots, a_K)$, with one coordinate per stopping process; the implemented model uses $\boldsymbol{a} = (a_1, a_2)$. Each a_j specifies the assumed dependence between the target event time T and stopping time C_j, conditional on the individual's features. For example, let T be time to finding employment and C_1 time to leaving follow-up. Setting $a_1 = 0.1$ assumes a weak positive relationship: among people with the same recorded features, earlier employment tends to accompany earlier departure from follow-up, and later employment tends to accompany later departure. The value specifies the correlation between the underlying Gaussian variables used to generate these times (Equation 3), rather than a 10% probability of leaving follow-up or a 10% change in predicted employment time.

The desired answer is an individual target-time distribution under that assumption and the study prior. DialPFN approximates this distribution with

$$q_{\theta,\phi}(T_* \mid \mathcal{D}, x_*, \boldsymbol{a}), \qquad S_{\boldsymbol{a}}(t) = P(T_* > t \mid \mathcal{D}, x_*, \boldsymbol{a}). \tag{2}$$

Here θ and ϕ are the backbone and quantile-head weights, respectively. T_* is the query individual's target event time and $t \geq 0$ is a chosen prediction horizon. The survival function $S_{\boldsymbol{a}}(t)$ is the probability that this target time exceeds t, conditional on the reference table, query features, and supplied assumption. Its left-hand notation suppresses the table and features for brevity.

The quantity $1 - S_{\boldsymbol{a}}(t)$ counts target times at or before t under the assumed joint-time model, including those hidden by an earlier stopping process. By contrast, recording the target event by t also requires it to occur before every stopping process. That probability is $P(T_* \leq t, T_* < \min_{1 \leq j \leq K} C_{j,*} \mid \mathcal{D}, x_*, \boldsymbol{a})$, where $C_{j,*}$ are the query individual's stopping times. For example, a simulated target at time 4 with observation ending at time 2 counts toward the first probability at $t = 5$, but not the second. Interpreting a latent target time as the outcome of preventing a competing event requires additional causal assumptions.

What a sensitivity curve means. For a fixed query and horizon, the curve is the function $a_j \mapsto S_{\boldsymbol{a}}(t)$ with the table and other coordinates fixed. Changing a_j changes the assumed relationship used to interpret incomplete records. The resulting curve shows how that individual's prediction depends on the assumption. A prediction interval gives a range of possible event times under one fixed assumption. Changing the assumption lets the analyst see how that range and the predicted probabilities change. Under any fixed assumption, the probability that the event has not yet occurred must decrease or stay unchanged as time passes. Changing the assumption is different: it can raise or lower this probability, depending on the observed records. A large change, or a change in direction as the assumption varies, therefore needs investigation before we treat it as a model error.

4 DialPFN: Learning and Prediction

4.1 Learning from Complete Simulated Studies

Each training task is a simulated study with its own event-time mechanisms and dependence setting. We construct its model inputs and training targets in six steps.

1. Sample study settings. We select a study family and sample the number of active features, their effects on event times, the parameters of the time distributions, and the dependence vector $\boldsymbol{a}$. These settings are shared by the individuals within a study and vary across studies.

2. Generate individual features. We draw between one and ten active features per individual from a standard normal distribution and zero-pad unused columns. The sampled feature effects determine how each individual's conditional time distributions differ.

3. Generate complete times. For the implemented $K = 2$ model, we draw the three times (T, C_1, C_2) for each individual through a Gaussian copula:

$$\begin{gathered}(Z_0, Z_1, Z_2) \sim \mathcal{N}(0, R_{\boldsymbol{a}}), \qquad R_{\boldsymbol{a}} = \begin{pmatrix} 1 & a_1 & a_2 \\ a_1 & 1 & 0 \\ a_2 & 0 & 1 \end{pmatrix}, \\ C_j = F_j^{-1}(\Phi(Z_j) \mid x), \qquad j = 0, 1, 2.\end{gathered} \tag{3}$$

Here F_0 is the target-time distribution and F_1, F_2 are the stopping-time distributions, conditional on the features and sampled study mechanisms. The simulator selects their distribution families, parameters, and feature effects before drawing individual times. These rules are shared within a simulated study and vary across studies; conditioning on x gives each individual their time distributions. For a real study, the analyst supplies the observed records, query features, and dependence assumption without specifying F_0, F_1, F_2.

The function Φ is the standard normal cumulative distribution function (CDF). It converts each Gaussian draw into its percentile rank: $U_j = \Phi(Z_j)$. For example, $Z_j = 0$ gives $U_j = 0.5$, the median. Because each Z_j is standard normal, each U_j is uniformly distributed between zero and one. The percentiles retain the Gaussian copula's dependence, although their ordinary correlations

can change. Applying $F_j^{-1}(U_j \mid x)$ then selects the time at that percentile in the individual's corresponding distribution. The implementation performs these calculations on log times, using log probabilities for the flexible mixture distributions and clipping extreme percentiles for the Weibull distributions to avoid numerical problems.

We cap each dependence coordinate at 0.9 to cover strong positive dependence while excluding near-perfect dependence. For the chosen correlation structure, a nonsingular Gaussian distribution requires $a_1^2 + a_2^2 < 1$. We impose the stricter bound $a_1^2 + a_2^2 \leq 0.92$ to maintain a numerical margin from this boundary. This guarantees a smallest correlation-matrix eigenvalue of at least $1 - \sqrt{0.92} \approx 0.0408$. The exact bounds are fixed training-design choices. Zero denotes pairwise conditional independence between the target and the corresponding stopping time in this construction; positive values associate earlier target times with earlier stopping times. Mixing features or studies can induce dependence even when a within-study coordinate is zero.

4. Construct observed records. We retain only the earliest time Y and its label J in each observed record. A stopping process that occurs first hides the target time from that record, while the simulator retains the complete target for supervision.

5. Form reference and query inputs. Disjoint rows from the same simulated study serve as reference and query records. The model receives features, observed times, and labels for the reference rows. Query inputs contain features and the stated dependence vector; their time and outcome-label inputs are masked. During training, the supplied dependence vector is the one used to generate the study.

6. Supervise hidden targets. Each query's complete simulated target time provides the training label, even when an earlier stopping event would hide it. These labels enter the training loss, while remaining hidden from the model's inputs.

Training includes studies with flexible event-time distributions and studies with Weibull-distributed event times. We vary feature effects and the proportions of observed outcomes, and include additional studies with few observed target events or weak dependence. We also swap the two stopping-process labels and their corresponding assumption coordinates so that training covers both label assignments. Appendix A specifies the distribution families, study mixture, and training schedule.

4.2 Conditioning Predictions on the Assumption

The network combines evidence from the reference records with the assumption supplied for the prediction. Each row embeds up to ten features plus the two assumption coordinates. A separate embedding encodes standardized observed log time and the three-way outcome label; query outcome inputs are zeroed. The transformer has eight attention blocks, width 256, and eight heads. Reference rows attend only to reference rows; a query attends to the reference rows and itself, not to other queries. Thus a query's prediction does not depend on which other query rows share its inference batch.

The backbone predicts a Gaussian mean μ_* and variance σ_*^2 for standardized log target time z_*. Its training loss is the Gaussian negative log-likelihood, averaged over training examples:

$$\mathcal{L}_{\mathrm{G}}(\theta) = \mathbb{E}\left[\tfrac{1}{2}\log(2\pi\sigma_*^2) + \frac{(z_* - \mu_*)^2}{2\sigma_*^2}\right]. \tag{4}$$

Standardization uses the reference table's observed log-time mean and spread. The Gaussian is defined over log time, so its implied physical times are positive. After synthetic training, the weights remain fixed: a new study changes the reference records, and a new assumption changes the input.

4.3 Constructing Event-Time Distributions

The backbone's Gaussian distribution limits the shapes it can represent. A small prediction component, called a quantile head, adjusts fifteen percentile values to allow more flexible distributions. We fit the head on simulated studies while keeping the backbone fixed.

Let h_* be the frozen 256-dimensional query representation and $g_k = \mu_* + \sigma_*\Phi^{-1}(p_k)$ its initial Gaussian quantiles, where σ_* is the predicted standard deviation and Φ^{-1} is the standard normal quantile function. The fixed probabilities satisfy $p_1 < \cdots < p_{15}$, with $p_8 = 0.5$. All quantiles use

the same standardized log-time scale as the target z_*. Writing $f_{\phi,k}(h_*)$ for the head's k-th output, we obtain

$$(\widetilde{q}_1, \ldots, \widetilde{q}_{15}) = \mathrm{sort}(g_k + 2\sigma_* \tanh f_{\phi,k}(h_*))_{k=1}^{15}. \tag{5}$$

The hyperbolic tangent $\tanh$ bounds each adjustment before sorting to within $2\sigma_*$ of its Gaussian quantile. Sorting orders the adjusted values from smallest to largest, producing $\widetilde{q}_k$ at level p_k. We fit these values using the pinball loss $\ell_p(u) = \max(pu, (p-1)u)$, where $u = z_* - \widetilde{q}_k$ is the target minus the predicted quantile:

$$\mathcal{L}_{\mathrm{Q}}(\phi) = \mathbb{E}\left[\frac{1}{15}\sum_{k=1}^{15} \ell_{p_k}(z_* - \widetilde{q}_k)\right]. \tag{6}$$

The sum averages over the fifteen quantiles; $\mathbb{E}$ averages over simulated training examples. The loss weights errors differently at each level p. The head has 35,343 parameters; together with the 4,287,746-parameter backbone, the predictor has 4,323,089 parameters. Both stages use simulated studies only. Appendix A gives the head architecture and quantile grid.

At prediction time, b_k is the quantile after retaining 75% of the learned adjustment, and $b_{\mathrm{med}} = b_8$ is the median. The final quantile q_k widens distances from this median by 5%:

$$b_k = g_k + 0.75(\widetilde{q}_k - g_k), \qquad q_k = b_{\mathrm{med}} + 1.05(b_k - b_{\mathrm{med}}). \tag{7}$$

These fixed development adjustments apply only during prediction, not training. Linear interpolation of (q_k, p_k) defines the CDF between quantile knots; the implementation assigns zero below the lowest knot and one above the highest. Coverage and the limits of this finite-tail treatment are evaluated separately from the training objective.

4.4 Averaging over Reference Samples

An individual's answer can vary with the records used as reference examples. To reduce dependence on one sample, we draw $B = 100$ reference tables $\mathcal{D}^{(b)}$ with replacement, each the same size as the original; b indexes the resample. Preprocessing is recalculated within each draw. Let $F_{\theta,\phi}$ denote the predicted target-time CDF in original time units, after undoing log-time standardization and exponentiating. It gives the predicted probability that the target occurs by t. We average these probabilities to obtain $\overline{F}$:

$$\overline{F}(t \mid x_*, \mathcal{D}, \boldsymbol{a}) = \frac{1}{B}\sum_{b=1}^{B} F_{\theta,\phi}(t \mid x_*, \mathcal{D}^{(b)}, \boldsymbol{a}). \tag{8}$$

The same reference draws are reused across assumptions. Consequently, differences along the sensitivity curve are not created by changing which random reference samples were used. Neither patients nor assumption settings are averaged together. A 90% prediction range is $[\overline{F}^{-1}(0.05), \overline{F}^{-1}(0.95)]$, not the average of component interval endpoints. Here $\overline{F}^{-1}$ is the quantile function of the averaged distribution.

Bagging preserves validity over time: a fixed nonnegative average of nondecreasing CDFs is nondecreasing and stays in $[0, 1]$. Its effect on accuracy remains an empirical question.

5 Evaluation Protocol

Seven checks assess accuracy under stated assumptions, preservation of prediction quality, and stability of sensitivity curves. Synthetic studies provide complete targets and known survival probabilities. Real datasets test ordinary prediction and sensitivity behavior; their hidden dependence is unknown. Appendix B gives the evaluation settings, with the checks' purposes and passing requirements in Sections B.3 and B.4, respectively.

Accuracy under stated assumptions. The *study-exit* and *competing-event benchmarks* compare the generating assumption with independence on the same studies, queries, and reference draws. We average study-level root mean squared error (RMSE) against known survival curves; confidence intervals use variation across studies. The *single-coordinate* and *joint-coordinate synthetic truth*

checks ask whether predictions favor inputs near the generating assumption, with one or both coordinates active. A reserved simulation family tests the frozen predictor beyond the training family.

Prediction quality and interval coverage. *Interval coverage preservation* checks how often complete simulated targets fall within the nominal 90% prediction range, compared with the earlier predictor. *Ordinary real-data prediction preservation* compares performance on 17 datasets using concordance, which measures event-time ranking, and integrated Brier score, which measures censoring-adjusted probability error over time. Higher concordance and lower Brier score are preferable.

Stability of individual sensitivity curves. *Population and individual stability* checks for abrupt changes and backtracking as the assumption varies, and for variation across reference samples. Fixed-query comparisons isolate the effect of changing reference records. The previously unused Melanoma dataset tests the frozen procedure separately. These checks use declared numerical limits rather than requiring predictions to move in one direction across assumptions.

Consistency across training runs. Training replication fixes six initializations and evaluates all six under one recipe. A preliminary weak-effect screen uses unbagged predictions; the subsequent seven-check suite uses 100 reference resamples.

6 Empirical Results

6.1 Accuracy under Stated Assumptions

Figure 1(a) asks whether supplying the correct assumption improves prediction. Both curves use the same model weights, observed records, and reference draws; only the assumption input changes. Lower error means predicted survival curves are closer to the known simulated truth. When the generating dependence is zero, both inputs are zero and the predictions coincide. At $\tau = 0.25$, supplying the generating assumption reduces mean individual RMSE from 0.04292 to 0.02682, a 37.5% reduction. At $\tau = 0.5$, it reduces RMSE from 0.07348 to 0.02719, a 63.0% reduction. The conclusion concerns the tested simulated worlds: supplying the correct dependence assumption improves accuracy over assuming independence.

The competing-event benchmark shows the same pattern (Table 1). All four confidence intervals for the paired error difference lie below zero, supporting lower mean error with the generating assumption. Strong-dependence reductions are 55.8% and 41.9% in the two banks. Appendix B.6 compares the survival PFNs. The generating assumption is known here because the studies are simulated.

The reserved studies use a dependence family absent from training. They test whether the predictor can use the assumption input under a different relationship between event times. Figure 2 holds each

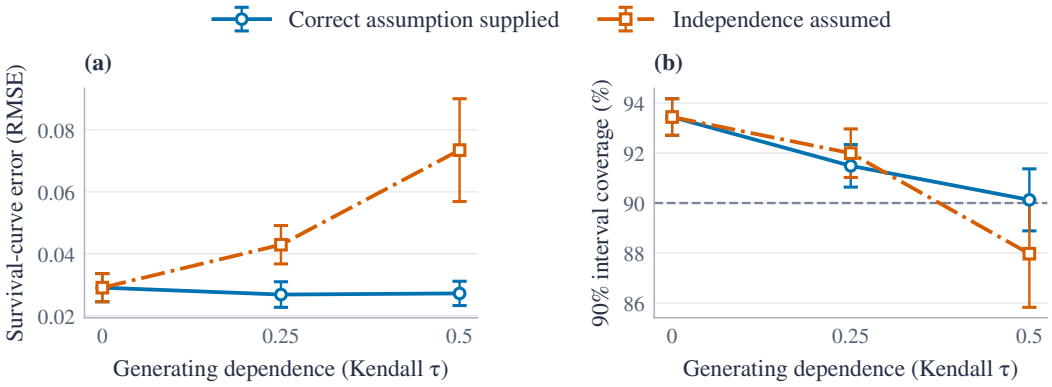

Figure 1: Reference predictor: (a) mean individual survival RMSE over 20 studies per setting; (b) mean interval coverage over 30 studies per setting. Bars show 95% confidence intervals across studies. The dashed line marks nominal 90% coverage.

Test bank	Dependence τ	Zero input	Generating input	Paired difference [95% CI]
1	0.25	0.0378	0.0271	-0.0107 [-0.0155, -0.0060]
1	0.5	0.0618	0.0273	-0.0345 [-0.0491, -0.0199]
2	0.25	0.0317	0.0256	-0.0061 [-0.0110, -0.0012]
2	0.5	0.0443	0.0257	-0.0185 [-0.0298, -0.0073]

Table 1: Competing-event benchmark: mean individual survival RMSE over 20 studies per bank and setting for the reference predictor. Paired differences subtract zero-input error from generating-input error; negative favors the generating input. CI denotes confidence interval.

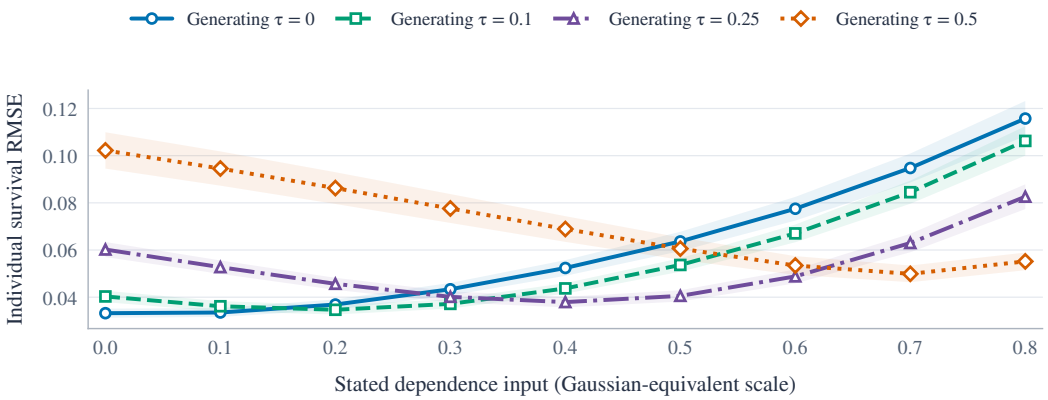

Figure 2: Reference predictor on reserved Frank-copula studies. Each curve fixes generating dependence and sweeps the stated input. Shading shows 95% confidence intervals across studies; lower is better.

generating setting fixed while varying the supplied assumption, keeping the observed records and model weights unchanged. Averaging errors over 100 studies per setting, the lowest error occurs at the tested input closest to the generating association on the Gaussian-equivalent scale, in all four settings. At strong dependence, the error reduction relative to zero is 48.3 to 54.7% across all five simulation banks. This supports useful assumption-conditioned prediction in this additional family, although it does not establish accuracy for arbitrary forms of dependence. These are repeated evaluations of one frozen predictor, not five training replications.

6.2 Prediction Quality and Interval Coverage

Accurate survival curves and reliable prediction ranges are separate goals. Figure 1(b) measures how often a nominal 90% range contains the complete simulated target time. With the generating assumption supplied, coverage is 93.44%, 91.49%, and 90.13% at $\tau = 0, 0.25, 0.5$, respectively. The corresponding 95% confidence intervals for mean coverage are [92.71,94.17], [90.64,92.34], and [88.89,91.37]%. Mean widths in log time are 2.55, 2.37, and 2.29. Coverage above 90% means that the ranges contain more targets than their nominal level requires. At strong dependence, the independence input covers 87.97%, with a 95% confidence interval of [85.83,90.12]%. Although this mean is below the target, its confidence interval still includes 90%. All settings pass the coverage-preservation rule, which compares with the earlier predictor rather than requiring exact 90% coverage.

Across 17 real datasets, mean concordance is 0.66747 compared with 0.66536 for the saved Gaussian-head baseline. Mean integrated Brier score is 0.16355 compared with 0.16462. These Brier means average the 80 valid dataset-and-division cells out of 85, using the same valid splits for both predictors. Datasets with more valid splits therefore receive more weight in this mean. Ordinary prediction is maintained, with little evidence of a substantial accuracy gain. Appendix C gives per-dataset scores. These values describe the reference predictor, rather than averages across training runs.

6.3 Stability of Individual Sensitivity Curves

We next vary the reference records while holding prediction patients fixed. Using the monoclonal gammopathy of undetermined significance dataset (MGUS2), we applied 100 reference resamples to the same 500 patients. Averaging each patient's distributions reduced the largest probability change between 0.1-spaced assumption settings to 7.94 percentage points. No patient at any tested time horizon exceeded 10 points. Each patient retains a separate prediction; averaging is over reference samples. In the broader development comparison, 25 and 50 samples also pass the grouped checks, but 10 fail MGUS2 stability. We keep the fixed count of 100 for all datasets.

Smoother curves could still be less accurate. A separate 32-study simulation comparison finds mean RMSE 0.028172 before averaging and 0.028120 after averaging, a small 0.18% reduction. The paired 95% confidence interval includes zero, so an overall accuracy improvement is not established. Nineteen studies improve and thirteen worsen. All eight strong-dependence studies improve, with a 2.61% reduction in mean error. Together with the full preservation checks, these results support averaging to reduce reference-sample sensitivity while maintaining useful prediction, without guaranteeing an accuracy gain on every study.

All five fixed divisions of the previously unused Melanoma dataset (205 records: 140 reference, 65 query per division) pass the population and patient stability checks. The largest individual change between 0.1-spaced settings is 1.73 percentage points. At any fixed setting and horizon, the largest gap between the five population averages is 5.02 points. The curves are stable on this previously unused dataset, but unknown hidden target times prevent checking their accuracy.

6.4 Consistency across Training Runs

Finally, repeating training tests whether the evaluated behavior extends beyond a single set of fitted weights. All six training runs pass every check. Section 5 introduces the checks. In Appendix B, Sections B.3 and B.4 explain their purposes and passing criteria, respectively. Passing requires more than smooth curves. In simulations, predictions must respond appropriately to zero, weak, and strong dependence, including when both coordinates are active. The other checks assess prediction quality, coverage, and stability.

Runs can pass the same checks and still differ in accuracy or coverage. The six-run summary therefore describes whether each run meets the stated requirements; it does not guarantee success for every future run. The numerical results above describe the reference predictor and should not be read as averages or confidence intervals across these training runs.

7 Scope and Limitations

Predictions remain conditional on the stated coordinate and training prior. We test bounded nonnegative dependence in a limited set of time-to-event mechanisms. Performance under negative association, additional stopping processes, arbitrary feature dimensionality, and other missing-target domains is not established here. Equal Kendall association does not remove copula misspecification.

The current output has only fifteen quantile knots and finite endpoint conventions. Long-horizon and extreme-tail claims require caution. Reference bagging costs repeated inference and does not provide a formal robustness or coverage guarantee. Stability limits are practical diagnostics; a steep response can reflect real assumption sensitivity, and a smooth curve can still be wrong. Ordinary real-data scores likewise depend on censoring assumptions and cannot validate unobserved outcomes.

8 Conclusion

DialPFN combines incomplete observed records with an explicit dependence assumption to produce an individual event-time distribution. Simulation supplies complete targets for learning this conditional prediction procedure, and in-context inference makes it reusable across studies. Holding the evidence fixed while varying the assumption exposes how predictions depend on relationships the observations alone cannot resolve.

ARTIFICIAL INTELLIGENCE USE STATEMENT

Generative artificial intelligence tools assisted with literature search and manuscript preparation, including drafting and polishing sections of the paper. The authors take responsibility for the content.

ETHICS STATEMENT

Making assumptions adjustable can expose how much a conclusion depends on them, but also lets an analyst choose settings that favor a preferred answer. Analysts should explain which assumptions they considered, why those settings are plausible, and how predictions change across them. If these predictions inform decisions about medical care or financial support, errors could harm the people affected. Such use requires validation in the intended setting, including checks for differences in accuracy and coverage across relevant population groups. We train on simulated studies and evaluate on existing research datasets. Reuse of those datasets should respect their access and redistribution terms and protect the privacy of the people represented. In particular, the simplified EBMT dataset is intended for illustration and cannot support clinical conclusions (Appendix C).

REPRODUCIBILITY STATEMENT

Appendix A describes the simulator, architecture, training settings, distribution construction, and reference resampling. Appendix B defines scoring and stability diagnostics. The figures and continuous-score tables are exported from saved results, without rerunning prediction or fitting. Exact source hashes and plotting scripts are retained with the local manuscript source. The source code supporting the findings of this study will be made publicly available upon official peer-reviewed acceptance of the manuscript.

REFERENCES

Tim Bedford. Assessing the impact of preventive maintenance based on censored data. *Quality and Reliability Engineering International*, 20(3):247–254, 2004.

On behalf of CRASH et al. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (crash-2): A randomised, placebo-controlled trial. *The Indian Journal of Neurotrauma*, 9(1):3–14, 2012.

Samuel Böhm, Lennart Purucker, Frank Hutter, and Pascal Schlosser. Survpfn: Towards foundation models for survival predictions. *arXiv preprint arXiv:2606.04564*, 2026.

Jordan M Braciszewski, Paul A Toro, and Robert L Stout. Understanding the attainment of stable housing: a seven-year longitudinal analysis of homeless adolescents. *Journal of Community Psychology*, 44(3):358–366, 2016.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Martha Fiocco, Hein Putter, and JC Van Houwelingen. Reduced rank proportional hazards model for competing risks. *Biostatistics*, 6(3):465–478, 2005.

Ronald B Geskus. On the inclusion of prevalent cases in hiv/aids natural history studies through a marker-based estimate of time since seroconversion. *Statistics in medicine*, 19(13):1753–1769, 2000.

T Rudra Gupta, DE Schwartz, Riddhiman Saha, PY Wen, Rifaquat Rahman, and Lorenzo Trippa. Informative censoring in externally controlled clinical trials: a potential source of bias. *ESMO open*, 10(1):104094, 2025.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23(1):77–91, 2004.

Emil Javurek, Dennis Frauen, Marie Brockschmidt, Jonas Schweisthal, and Stefan Feuerriegel. Amortizing causal sensitivity analysis via prior data-fitted networks, 2026.

David Koehler. Crash 2 competing risk data. R documentation, discSurv package, n.d. URL https://cran.r-project.org/web/packages/discSurv/refman/discSurv.html#crash2. Documentation for package version 2.5.1. Accessed 2026-09-25.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Janice R Offord, Dirk R Larson, Matthew F Plevak, and L Joseph Melton III. A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 346(8):564–569, 2002.

Olof Liberg, Guillaume Chapron, Petter Wabakken, Hans Christian Pedersen, N Thompson Hobbs, and Håkan Sand. Shoot, shovel and shut up: cryptic poaching slows restoration of a large carnivore in europe. *Proceedings of the Royal Society B: Biological Sciences*, 279(1730):910, 2011.

Si-Yang Liu and Han-Jia Ye. Tabpfn unleashed: A scalable and effective solution to tabular classification problems. 2025.

mstate AIDSSI documentation. *Data from the Amsterdam Cohort Studies on HIV Infection and AIDS*. mstate R package, n.d. URL https://cran.r-project.org/web/packages/mstate/refman/mstate.html#aidssi. R documentation, aidssi; package version 0.3.3. Accessed 2026-09-25.

mstate EBMT documentation. *Data from the European Society for Blood and Marrow Transplantation (EBMT)*. mstate R package, n.d. URL https://cran.r-project.org/web/packages/mstate/refman/mstate.html#EBMT+20cause+20of+20death+20data. R documentation, EBMT cause of death data; package version 0.3.3. Accessed 2026-09-25.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. 2021.

National Cancer Institute. SEER cause-specific death classification. SEER data documentation, n.d. URL https://seer.cancer.gov/causespecific/. Accessed 2026-09-25.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. volume 29, 2016.

Hein Putter, Marta Fiocco, and Ronald B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.

Shi-ang Qi, Vahid Balazadeh, Michael Cooper, Russell Greiner, and Rahul G Krishnan. Survivalpfn: Amortizing survival prediction via in-context bayesian inference. 2026.

survival MGUS2 documentation. *Monoclonal Gammopathy Data*. survival R package, n.d. URL https://cran.r-project.org/web/packages/survival/refman/survival.html#mgus2. R documentation, mgus2; package version 3.8-12. Accessed 2026-09-26.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

Gerard J Van Den Berg and Maarten Lindeboom. Attrition in panel survey data and the estimation of multi-state labor market models. *Journal of Human Resources*, pp. 458–478, 1998.

Chih-Tung Yeh, Gen-Yih Liao, and Takeshi Emura. Sensitivity analysis for survival prognostic prediction with gene selection: a copula method for dependent censoring. *Biomedicines*, 11(3): 797, 2023.

A IMPLEMENTATION AND TRAINING DETAILS

A.1 STUDY PRIOR

All implementation details below describe the evaluated $K = 2$ model. The base study generator samples between one and ten active features and zero-pads unused feature columns. Features are standard normal within each simulated study. Target log time includes a sampled intercept, sparse linear effects, and an optional two-feature interaction. Stopping-time margins have their own sampled location and scale, with feature effects in one stopping process. Flexible margins apply monotone mixture-based transforms to Gaussian-copula uniforms; the complementary Weibull family uses an inverse cumulative distribution function (CDF) sampling procedure for a Weibull distribution. The observed log time is the minimum of the complete log times, and the minimum's index gives a one-hot outcome label.

Training uses a shuffled schedule with 200,000 base-pool batches, 50,000 expanded-pool batches, and 50,000 targeted continuous-pool batches. The expanded pool includes flexible and Weibull families and varies the fraction of records whose observation ends through a competing label. The targeted pool places probability 0.4 on a sparse-target configuration, 0.3 on balanced outcomes, and 0.3 on a broader configuration. The sparse configuration accepts reference tables with target-event share 0.05 to 0.15, one stopping-label share 0.55 to 0.75, and the other share 0.15 to 0.35 before optional label swapping. Selection uses observed reference labels, not hidden query outcomes.

The targeted dependence sampler allocates equal probability to four strata: first coordinate zero and second in $[0, 0.3]$; second zero and first in $[0, 0.3]$; both in $[0, 0.3]$; and the full valid region of Equation 3. The same sampled dependence applies within a study. Swapping the stopping labels also swaps their coordinates. These choices increase representation of weak association and imbalanced observed tables; they do not impose directionality on predictions conditional on those tables.

A.2 TRAINING SETTINGS AND PARAMETER COUNT

The architecture uses Gaussian error linear units (GELU) and layer normalization (LayerNorm). The optimizer is adaptive moment estimation with decoupled weight decay (AdamW).

Component	Configuration
Backbone	8 layers, width 256, 8 attention heads
Feed-forward blocks	Width 512, GELU, residual connections
Backbone updates	300,000; final checkpoint used
Batch	64 simulated studies; 80 query rows per study
Reference rows during backbone training	60 to 600
Backbone optimizer	AdamW; initial rate 3×10^{-4}; cosine schedule
Weight decay	10^{-4}
Quantile head	LayerNorm, 256-to-128, GELU, 128-to-15
Head fitting	2,000 simulated studies; 64 query rows per study
Head validation studies	400 (reference predictor); 1,000 (replication)
Reference rows during head fitting	100 to 400
Head optimizer	AdamW; rate 10^{-3}; batch 2,048 query representations
Head stopping	At most 60 epochs; patience 8; minimum improvement 10^{-5}
Inference	Correction 0.75; width factor 1.05; 100 reference resamples

Table 2: Architecture and training settings for the reference predictor and replication runs.

The 35,343 head parameters are explicit: LayerNorm contributes $2 \times 256 = 512$; the first linear layer contributes $256 \times 128 + 128 = 32,896$; the output layer contributes $128 \times 15 + 15 = 1,935$. Their sum is 35,343. The final linear layer starts at zero, so the initial head reproduces the Gaussian quantile scaffold. The backbone remains frozen while fitting the head. Selection among head epochs uses synthetic validation loss, not real-data outcomes.

The fifteen probabilities are

$$\{0.025, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.95, 0.975\}.$$

These are positions within a predicted event-time distribution. They are not dependence settings. For example, the 0.5 quantile is a median target time under one particular assumption and reference table.

A.3 Reference averaging procedure

For each observed table, construct 100 resamples of n rows with replacement. Duplicates are retained. The sequence of draws is determined from reference records, not query outcomes or the stated coordinate. All queries and all settings in a sweep use this same sequence. For each draw, recompute the reference-based feature and time transformations, obtain query quantiles, undo log-time standardization, and construct a distribution on the common scale. Average these distributions as in Equation 8.

For illustration, 600 draws with replacement from 600 reference records typically include fewer than 600 distinct records. This is intentional bootstrap resampling, not a smaller fixed context shared by every repeat. The query record never becomes part of the reference sample. Its prediction can vary because the records used as examples vary, even though its own features, the model weights, and the stated assumption stay fixed.

The average of valid distributions remains valid, but marginal event-time prediction intervals and reference-resampling intervals answer different questions. We report the former by inverting the mixture CDF. We do not label the spread of component medians as a 90% event-time prediction range. Likewise, neither the 100 components nor the prediction patients constitute 100 independent model-training replications.

B Evaluation definitions and scope

B.1 Sampling units and scoring

The study-exit benchmark has 20 studies at each generating Kendall rank association $\tau \in \{0, 0.25, 0.5\}$, with 500 reference and 500 query records per study. The competing-event benchmark uses two fixed banks of 20 studies at $\tau = 0.25$ and 0.5. Comparisons reuse the same studies, queries, and reference draws for generating-assumption and independence inputs. For other copula families, we map τ to the Gaussian-equivalent input $a = \sin(\pi\tau/2)$. This aligns rank association while leaving differences in dependence structure and tails.

A training seed, simulation-bank seed, patient division, and bootstrap draw refer to four different sources of variation. The 100 reference resamples are components of each prediction, not independent evaluation studies.

For a study with N query records and L evaluation times, individual survival root mean squared error (RMSE) is

$$\mathrm{RMSE}_{\mathrm{indiv}} = \sqrt{\frac{1}{NL}\sum_{i=1}^{N}\sum_{\ell=1}^{L}\left(\widehat{S}_i(t_\ell) - S_i^{\mathrm{true}}(t_\ell)\right)^2}. \tag{9}$$

We average this study-level score across studies. Population RMSE first averages survival probabilities over query records and is a different quantity. In the main study-exit benchmark, ten evaluation times span the 10th to 90th percentiles of reference observed log times. Confidence intervals are computed from study-level scores with a Student-t multiplier. A paired difference uses the difference of scores within each study before calculating its interval. It is not obtained by subtracting the endpoints of two separate intervals.

Coverage uses 30 studies per dependence setting, each with 500 independently sampled query target times from the known conditional target margin. Each study contributes a fraction covered by its nominal 90% prediction range and a mean prediction width. Averaging these fractions yields the reported coverage; the uncertainty interval uses variation across the 30 studies. We report width on the log-time scale. Ordinary prediction uses five reference/query divisions of each of 17 real datasets; dataset-specific scoring details appear in Appendix C. The null setting, where truth and zero inputs are identical, is a consistency check rather than evidence of a benefit from changing assumptions.

The broader single-coordinate suite uses five fixed simulation banks with 20 studies per dependence setting and 500 reference plus 500 query records. Its generating associations are $\tau \in \{0, 0.1, 0.25, 0.5\}$ and the input sweep is $\{0, 0.1, \ldots, 0.8\}$. The joint-coordinate test uses flexible and Weibull three-outcome worlds at generating pairs $(0, 0), (0.4, 0.4), (0.8, 0.3), (0.3, 0.8)$. Each pair and family has ten studies per simulation bank. One coordinate is swept while the other is held at truth; pinball loss scores the predicted quantiles against complete query targets. A good marginal sweep alone would not establish this joint behavior.

B.2 Numerical stability diagnostics

The coarse assumption sweep is $0, 0.1, \ldots, 0.8$. Let $s_g = S_{\boldsymbol{a}^{(g)}}(t)$, $g = 0, \ldots, G$, denote one fixed patient's probability at one fixed horizon along a grid of $G+1$ assumption settings. One dependence coordinate varies along the grid while the others remain fixed. Define

$$L(s) = 100 \max_{0 \le g < G} |s_{g+1} - s_g|, \qquad E(s) = 100 \left(\sum_{g=0}^{G-1} |s_{g+1} - s_g| - |s_G - s_0| \right). \tag{10}$$

Both are in percentage points. L measures the largest local change. E is zero for any monotone sequence in either direction, and positive when a path moves back and forth. It does not assert that every positive value is scientifically wrong. The reported flag rate is the fraction of patient-and-horizon combinations exceeding a limit, not the fraction of distinct patients unless only one horizon is used.

Population diagnostics apply the same functions to the mean probability at each assumption. Repeatability compares population probabilities across fixed reference/query divisions. Because a division can change both sets of patients, this quantity is not a pure estimate of sensitivity to reference selection alone. The fixed-query reference-bootstrap diagnosis isolates that part more directly. Population smoothness cannot establish patient-level smoothness: different individuals can make large changes at different settings that are small after averaging.

The coarse-grid development budgets are $L \le 10$ and $E \le 10$ for population curves, at most 5% of patient-horizon paths above each 10-point budget, and population spread across divisions no greater than 10 points. Probability validity requires values in $[0, 1]$ and survival nonincreasing over supplied time horizons. The additional dense MGUS2 population grid uses spacing 0.025. These numerical budgets were developed on exposed data and are not field-wide standards or clinical utility thresholds. They are not untouched evidence in themselves. Fixed-query bootstrap comparisons isolate changes in reference composition; the sample-count comparison evaluates 10, 25, 50, and 100 resamples. The separate confirmation uses a frozen predictor and procedure.

B.3 What the seven groups protect

The study-exit benchmark protects useful prediction under a stated target/exit association. The competing-event benchmark checks the same mechanism under a different labeled stopping process. Coverage protects prediction ranges against changes from the earlier model. Ordinary real-data prediction protects ranking and probability fit under standard scoring assumptions. Stability checks numerical behavior of both individual and population sensitivity paths. Single-coordinate synthetic truth tests null honesty and preference near the generating association. Joint-coordinate truth verifies that one input retains useful meaning while another is active. A grouped pass therefore summarizes multiple conditions, not a single scalar score.

The weak-effect screen is deliberately separate from the full bagged suite. It tests the trained head without reference averaging before committing to the much larger evaluation. Results from an unbagged screen should not be confused with a measured result for the 100-sample predictor. The six-run summary reports successful completion of both stages.

B.4 Numerical decision rules

For error comparisons, let Δ denote the paired study-level score under the generating assumption minus the score at zero. Negative values favor the generating assumption. A confidence bound below refers to the study-level paired 95% interval, not uncertainty across training runs. The following are

engineering preservation and response criteria, not field-wide definitions of an acceptable survival model.

Study-exit benchmark. Both individual and population mean RMSE may worsen by at most 3% relative to the saved original version 1 Gaussian-head predictor in each tested dependence setting. Generating-input and zero-input individual errors must agree at the null within 10^{-12}; at nonzero settings the generating input must have strictly lower mean individual RMSE.

Competing-event benchmark. At mild dependence, require $\Delta \leq 0$ and its upper confidence bound at most 0.002. At strong dependence, require at least 15% error reduction relative to zero and lower mean error than both saved SurvivalPFN and SurvPFN benchmark scores. Both simulation banks must satisfy these rules. These saved comparisons concern the specified benchmark configurations, not a claim of superiority over every implementation or tuning regime.

Coverage. The absolute change in 90% interval coverage from the earlier predictor must be at most 0.03 in every tested setting. Passing does not mean coverage equals 0.90.

Ordinary prediction. Mean concordance may decrease by at most 0.003 and mean integrated Brier score may increase by at most 0.003 relative to the earlier predictor.

Stability. Apply the population, patient, and repeatability budgets defined above, together with probability validity. No fixed direction across assumptions is required.

Single-coordinate truth. For competing-event sweeps, at the null the best input must be at most 0.2, the zero-input error minus the best error at most 0.002, and the lower paired bound for best minus zero at least -0.002. At weak dependence, the best input must lie within 0.1 of the nearest swept generating input. At moderate dependence, require $\Delta \leq 0$ and upper bound at most 0.002. At strong dependence, require at least 15% error reduction in every simulation bank. A separate study-exit guard requires bagged quantile mean individual RMSE at the nearest swept generating input to worsen by at most 3% against the same backbone's original unbagged Gaussian output, in every bank and dependence setting.

Joint-coordinate truth. Use quantile loss. At the null, use the same 0.2 input and 0.002 loss tolerances as above. At nonzero generating pairs, the best swept input must lie within 0.2 of the nearest generating input; the generating-input loss difference from zero must be nonpositive with upper bound at most 0.002. Apply these rules to both coordinate sweeps and both world families.

The preliminary weak-effect screen uses the unbagged corrected quantile head at generating Kendall association $\tau = 0.1$. The best input must be one of $\{0.1, 0.2, 0.3\}$ in both the Gaussian and Clayton families. This screen checks a local response before the seven full groups; it does not replace any of them. The Gaussian-equivalent input for this generating association is approximately 0.156, whose nearest swept value is 0.2. In the single- and joint-coordinate truth suites, comparisons at the generating input use its nearest value on the specified sweep grid.

B.5 Reserved confirmation

The reference predictor is frozen before confirmation in a reserved simulation family using a Frank copula, a log-normal target margin, and a Weibull competing-time margin. This tests conditional prediction beyond the training family. The previously unused Melanoma dataset provides a separate behavioral check after the prediction procedure is frozen. Its hidden dependence remains unknown, so this real-data check assesses stability rather than correctness of a stated assumption.

B.6 Comparison with survival PFNs

Table 3 compares the reference predictor with saved SurvivalPFN (Qi et al., 2026) and SurvPFN (Böhm et al., 2026) scores on the same competing-event studies. Each cell averages 20 studies with 500 reference and 500 query records; banks 1 and 2 use simulation seeds 0 and 2026. All methods are scored against the known individual target-survival curves on the same time grid.

	Bank 1		Bank 2	
Predictor and input	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.25$	$\tau = 0.5$
SurvivalPFN	0.0722	0.1030	0.0814	0.0918
SurvPFN	0.0868	0.1129	0.0888	0.1006
DialPFN, zero	0.0378	0.0618	0.0317	0.0443
DialPFN, generating	0.0271	0.0273	0.0256	0.0257

Table 3: Mean individual survival RMSE on the competing-event benchmark. Lower is better.

Both survival PFNs treat the competing event as censoring. DialPFN uses 100-reference-sample averaging and is shown with both zero and generating assumption inputs. The generating input supplies information that the baselines do not receive. DialPFN has lower mean error in all four settings even at zero input; supplying the generating assumption reduces it further. These are descriptive comparisons of the saved configurations, rather than a tuning study or a test that isolates architecture. SurvivalPFN receives shifted log times, while SurvPFN receives times on the original scale, as implemented in the saved benchmark. Paired confidence intervals in Table 1 compare DialPFN's two inputs, not DialPFN against these baselines.

C Real-data prediction and interpretation

Dataset labels in the table retain their source identifiers; GBSG2 denotes the German Breast Cancer Study Group 2 dataset. For ordinary prediction, each division uses up to 400 reference records and

Dataset	Concordance ↑	Integrated Brier ↓	Brier cells
Framingham	0.6638	0.1439	5/5
GBSG2	0.6668	0.1913	5/5
LeukSurv	0.6820	0.1825	4/5
Pbc3	0.7763	0.0839	5/5
UnempDur	0.5383	0.2277	5/5
actg	0.7650	0.0603	5/5
chop	0.5795	0.2199	5/5
cost	0.6760	0.1796	5/5
dataDIVAT3	0.6547	0.0457	5/5
dataOvarian1	0.5943	0.1996	5/5
flchain	0.8051	0.1119	5/5
grace	0.6911	0.1813	5/5
phpl04K8a	0.6319	0.2070	5/5
prostate	0.5374	0.2188	5/5
rhc	0.6332	0.2206	2/5
rott2	0.7055	0.1765	5/5
whas500	0.7461	0.1690	4/5

Table 4: Reference-predictor scores averaged over five reference/query splits per dataset. Brier means exclude invalid splits; the final column reports the number of valid splits.

300 query records, with reference size capped at 60% of the dataset. The risk score for concordance is the negative predicted median log time. Integrated Brier scoring uses twelve times between the query 10th percentile and the minimum of the query 80th percentile and 99% of the reference and query maximum times. Query outcomes determine this evaluation window only; they are not inputs to the predictor. Censoring weights are estimated from reference outcomes. Five cells have no valid Brier score and remain absent from both compared predictors' means.

Real sensitivity datasets serve different purposes from this prediction table. GBSG2 occurs in both: here it measures ordinary prediction, while a separate fixed-table sweep examines sensitivity to study exit. The Amsterdam Cohort Studies dataset (AIDSSI) records the first of acquired immunodeficiency syndrome (AIDS) diagnosis and syncytium-inducing (SI) phenotype appearance. The cohort is described by Geskus (2000); the competing-event example follows Putter et al. (2007) and is distributed as `aidssi` in the mstate package (mstate AIDSSI documentation). The European

Society for Blood and Marrow Transplantation (EBMT) cause-of-death data are associated with Fiocco et al. (2005). We use the simplified `ebmt2` version distributed by mstate for illustration; its documentation explicitly rules out clinical conclusions (mstate EBMT documentation). The Clinical Randomisation of an Antifibrinolytic in Significant Haemorrhage trial (CRASH2) is reported by behalf of CRASH et al. (2012). We use the adapted competing-event data supplied with the discSurv package (Koehler, n.d.). MGUS2 comes from the follow-up study of Kyle et al. (2002); the survival package's `mgus2` data provide progression, death, and follow-up information (survival MGUS2 documentation). The previously unused Melanoma dataset from the R package accompanying *Modern Applied Statistics with S* (MASS) is reserved for the separate behavioral confirmation, not for choosing training or bagging settings.